ACOUSTIC SCENE CLASSIFICATION USING DEEP NEURAL NETWORK AND FRAME-CONCATENATED ACOUSTIC FEATURE

Gen Takahashi[1], *Takeshi Yamada*[1], *Shoji Makino*[1] *and Nobutaka Ono*[2]

[1]University of Tsukuba, Japan
[2]National Institute of Informatics / SOKENDAI, Japan
g.takahashi@mmlab.cs.tsukuba.ac.jp

ABSTRACT

This paper describes our contribution to the task of acoustic scene classification in the DCASE2016 (Detection and Classification of Acoustic Scenes and Events 2016) Challenge set by IEEE AASP. In this work, we applied the DNN-GMM (Deep Neural Network-Gaussian Mixture Model) to acoustic scene classification. We introduced high-dimensional features that are concatenated with acoustic features in temporally adjacent frames. As a result, it was confirmed that the classification accuracy of the DNN-GMM was improved by 5.0% in comparison with that of the GMM, which was used as the baseline classifier.

Index Terms— acoustic scene classification, DNN, MFCC, frame concatenation

1. INTRODUCTION

Recently, a DNN (Deep Neural Network) that has a multilayer neural network has been actively investigated. In general, a DNN tends to fall into a local solution and requires an unrealistic learning time. However, a pre-training method that gives appropriate initial values and high-speed computation on a GPU (Graphics Processing Unit) have been established. Because of this, a DNN is now being applied in various classification problems.

For speech recognition, a DNN-HMM, which combines a DNN and an HMM (Hidden Markov Model), has been proposed [1]. The probability distribution in an HMM is generally represented by a GMM (Gaussian Mixture Model). On the other hand, it is precisely represented by the DNN in the DNN-HMM. It was reported that the performance of speech recognition is markedly improved by using the DNN-HMM [2].

In this work, we applied the DNN-GMM to the task of acoustic scene classification, in the DCASE2016 Challenge and evaluated its performance. The features used for classification are the MFCC (Mel Frequency Cepstral Coefficient) along with its first and second differences. Features in temporally adjacent frames are concatenated to form high-dimensional acoustic features. We perform acoustic scene classification by inputting these features into the DNN-GMM.

2. SYSTEM OVERVIEW

2.1. Process flow

Figure 1 shows the process flow of our system. First, we compute acoustic features (the MFCC and its first and second differences) in each frame for each of the left and right channels. Next, we concatenate acoustic features in each frame with those in several frames

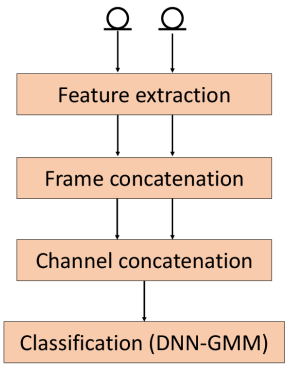

Figure 1: Process flow

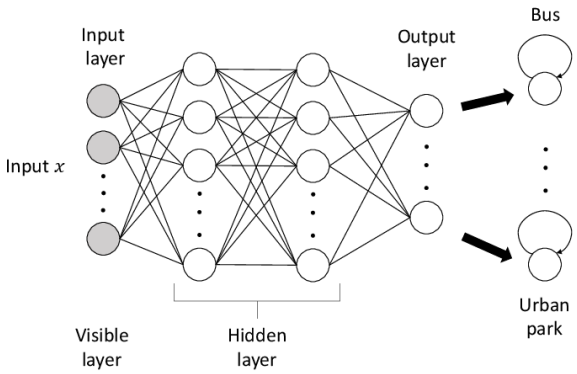

Figure 2: Example of the DNN-GMM

before and after the frame for each of the left and right channels. We then concatenate the acoustic features of the left and right channels in each frame. Finally, we perform acoustic scene classification by inputting the high-dimensional acoustic features into the DNN-GMM.

2.2. DNN-GMM

Figure 2 shows an example of the DNN-GMM. The DNN-GMM is basically the same as the DNN-HMM but with the GMM (one-state HMM without state transition) used instead of the HMM. Each GMM corresponds to one of the acoustic scenes.

The DNN training is divided into unsupervised pre-training to

Table 1: Overview of the development dataset

# of classes	15
# of sound data	1170 (=15 classes×78 data)
data length	30 s
# of channels	2 (left and right)
sampling frequency	44.1 kHz
quantization bits	16 bits

Table 2: Conditions of the acoustic features and the DNN-GMM

feature	MFCC+Δ + $\Delta\Delta$ (60 dimensions)×n frames
frame concatenation n	1, 3, 5
hidden layer	2, 3, 4, 5
dimension of hidden layer	128, 256, 512, 1024, 2048

obtain the appropriate initial value and supervised fine-tuning. In this work, we first perform the pre-training processing using the CD (Contrastive Divergence) method [3] by regarding each layer as an RBM (Restricted Boltzmann Machine). Next, we add the softmax layer initialized using a random number and perform backpropagation using the SGD (Stochastic Gradient Descent) method.

3. EVALUATION

3.1. Experimental conditions

In this experiment, we evaluated the performance of the DNN-GMM by using the development dataset provided by the DCASE2016 Challenge. Table 1 gives an overview of this dataset. The dataset contains 15 classes of acoustic scenes. Each class has 78 sound data, each of which is a stereo signal with a duration of 30 s. The sampling frequency is 44.1 kHz and the number of quantization bits is 16 bits.

Table 2 shows the conditions of the acoustic features and the DNN-GMM. The acoustic features used are the 20th-order MFCC and its first and second differences. The frame length and frame rate in the frame analysis are 40 ms and 20 ms, respectively. By the concatenation of n frames and the concatenation of two channels, the dimension of features becomes $20 \times 3 \times n\ (\mathrm{frame}) \times 2\ (\mathrm{ch})$. The frames used for frame concatenation are selected at 100 ms (5 frame) intervals. In this experiment, the number of frame concatenations n is set to 1, 3 or 5.

The number of hidden layers of the DNN (excluding the input layer and softmax layer) is set to 2, 3, 4 or 5, and the dimension of each hidden layer is constant and sets to 128, 256, 512, 1024 or 2048. We performed the pre-training processing using the CD-1 method on the RBM of the hidden layer and using the CD-2 method on the RBM of the input layer. We set the learning rate of the RBM to 0.4, the learning rate of the DNN to 0.008 and the dropout rate to 0.0 on the basis of the results of a preliminary experiment. We performed four-fold cross-validation on the development dataset. We used Kaldi [4] to build our system.

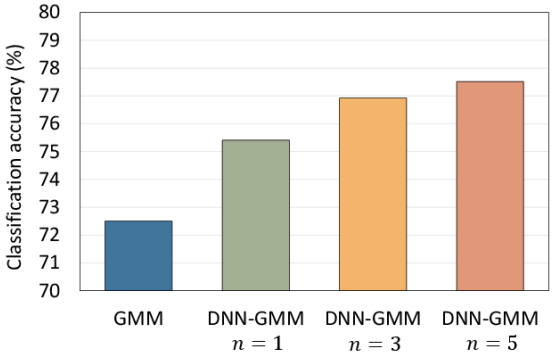

Figure 3: Classification accuracy

3.2. Results

Figure 3 shows the results of the experiment. The vertical axis of this figure shows the classification accuracy (the average of the results obtained from four-fold cross-validation). The classification accuracy of the DNN-GMM was improved by 5.0% compared with that of the GMM. We can see that the improvement in performance increases with n. The highest classification accuracy is 77.5%, which was achieved by the DNN-GMM ($n = 5$) with three hidden layers and 2048 dimensions in each hidden layer.

Furthermore, we also evaluated the performance of the DNN-GMM by using the evaluation dataset. The dataset has 390 sound data for testing, which are different from those in the development dataset. We used all the sound data in the development dataset for training the DNN-GMM ($n = 5$) with three hidden layers and 2048 dimensions in each hidden layer. The other conditions were the same as in the experiment mentioned above. The classification accuracy of the DNN-GMM was 85.6%, which was improved by 8.4% compared with that of the GMM.

4. CONCLUSION

In this work, we applied the DNN-GMM to the task of acoustic scene classification in the DCASE2016 Challenge and evaluated its effectiveness. It was confirmed that the classification accuracy of the DNN-GMM was improved by 5.0% and 8.4% in comparison with that of the GMM on the development dataset and the evaluation dataset, respectively, and that frame concatenation is particularly effective.

5. REFERENCES

[1] H.A. Bourlard, N. Morgan, "Connectionist speech recognition: a hybrid approach," Vol. 247, Springer, 1994.

[2] G.E. Hinton, L. Deng, Y. Dong, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, Vol. 29, No. 6, pp. 82-97, 2012.

[3] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, Vol. 14(8), pp. 1771-1800, 2002.

[4] `http://kaldi-asr.org/`.